Journal of Zhejiang University-SCIENCE A 2025 26(10):895-916
www.jzus.zju.edu.cn; www.springer.com/journal/11582
E-mail: jzus_a@zju.edu.cn

Research Article
https://doi.org/10.1631/jzus.A2400560

Check for updates

Virtual sample diffusion generation method guided by large language model-generated knowledge for enhancing information completeness and zero-shot fault diagnosis in building thermal systems

Zhe SUN[1], Qiwei YAO[1], Ling SHI[1], Huaqiang JIN[3✉], Yingjie XU[1], Peng YANG[1], Han XIAO[1], Dongyu CHEN[4], Panpan ZHAO[5], Xi SHEN[1,2✉]

[1]*College of Mechanical Engineering, Zhejiang University of Technology, Hangzhou 310023, China*
[2]*College of Information Science and Engineering, Jiaxing University, Jiaxing 314001, China*
[3]*College of Education, Zhejiang University of Technology, Hangzhou 310023, China*
[4]*College of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*
[5]*Hefei General Machinery Research Institute Company Limited, Hefei 230031, China*

Abstract: In the era of big data, data-driven technologies are increasingly leveraged by industry to facilitate autonomous learning and intelligent decision-making. However, the challenge of "small samples in big data" emerges when datasets lack the comprehensive information necessary for addressing complex scenarios, which hampers adaptability. Thus, enhancing data completeness is essential. Knowledge-guided virtual sample generation transforms domain knowledge into extensive virtual datasets, thereby reducing dependence on limited real samples and enabling zero-sample fault diagnosis. This study used building air conditioning systems as a case study. We innovatively used the large language model (LLM) to acquire domain knowledge for sample generation, significantly lowering knowledge acquisition costs and establishing a generalized framework for knowledge acquisition in engineering applications. This acquired knowledge guided the design of diffusion boundaries in mega-trend diffusion (MTD), while the Monte Carlo method was used to sample within the diffusion function to create information-rich virtual samples. Additionally, a noise-adding technique was introduced to enhance the information entropy of these samples, thereby improving the robustness of neural networks trained with them. Experimental results showed that training the diagnostic model exclusively with virtual samples achieved an accuracy of 72.80%, significantly surpassing traditional small-sample supervised learning in terms of generalization. This underscores the quality and completeness of the generated virtual samples.

Key words: Information completeness; Large language models (LLMs); Virtual sample generation; Knowledge-guided; Building air conditioning systems

1 Introduction

1.1 Background

In the big data era, vast amounts of data are generated daily across industries, driving advances in data mining and machine learning. However, big data's value lies in its information, not its volume. Repetitive and homogeneous data adds little value and burdens processing tasks. This abundance of data but lack of valuable knowledge has led to the "small samples in big data" problem (Qi and Luo, 2022). This issue refers to data samples with limited information content, indicating their incompleteness.

In China, newly constructed buildings with air conditioning systems typically have sensing devices and data storage capabilities, generating vast amounts of operational data (Li et al., 2021, 2023b; Liu et al., 2023). However, these systems operate in complex, dynamic environments influenced by seasons, weather, locations, and user habits, causing significant variability in system characteristics (Zhang et al., 2024b). To understand these dynamics with machine learning, data

✉ Xi SHEN, sx@zjut.edu.cn
Huaqiang JIN, Jhq@zjut.edu.cn
Zhe SUN, https://orcid.org/0000-0002-9664-9749

Received Dec. 5, 2024; Revision accepted Mar. 17, 2025; Crosschecked Sept. 16, 2025

obtained under various conditions are essential. In reality, most data collected are from similar conditions, resulting in incomplete information. This is especially problematic for fault diagnosis, where fault samples are scarce, and only abundant healthy samples under uniform conditions are available (Guo et al., 2024). Training models with such data hamper their ability to adapt to complex, changing situations, leading to poor generalization, a significant issue for data-driven methods.

There are two main approaches to address the problem of incomplete information. The first is optimizing the model to improve learning efficiency and extrapolation with limited data, such as adding regularization to neural networks or using transfer learning from related fields (Li et al., 2023a). This approach generally only partially alleviates the issue. The second approach focuses on optimizing the data by increasing the information content of the training data, thereby fundamentally resolving the problem. Virtual sample generation is a method representative of this approach. Virtual samples, also known as synthetic or artificial samples, are created to contain valid information (Tian et al., 2023). This helps the estimated hyperplane $\hat{H}$ of the small sample set align more closely with the overall hyperplane H, enhancing the completeness of the information.

Virtual sample generation includes three main methods: basic transformations or resampling, machine learning, and information diffusion. Basic transformations or resampling involves creating similar samples from the original ones using techniques such as rotation, scaling, or interpolation. This method is common in image processing, where datasets are augmented using techniques like rotation, lighting adjustments, and background replacement (Xu ML et al., 2023). Although it can enhance sample information, its effectiveness is limited. For instance, interpolating between two samples often results in a smooth line connecting the sample points, making it difficult to capture inflection points.

Generation methods based on machine learning are more effective than the first approach. Through data learning, they can achieve generation performance that surpasses smooth interpolation based on probabilistic reasoning. This method has advantages for generative modeling of complex physical fields, such as vortex flow (Li L et al., 2024b, 2025a) and multiphase flows (Tan et al., 2023; Li Z et al., 2024; Li L et al., 2025b; Xu et al., 2025), where mechanistic generative modeling is very difficult. Sanchez et al. (2023) used GAN (generative adversarial network) for generative modeling of two-phase flow imaging, and Du et al. (2024) proposed a VAE (variational autoencoder)-GAN-based generative modeling method for behavior recognition in bubbly flow. Whittaker et al. (2024) proposed a diffusion model-based generative modeling method for the turbulence scaling problem. However, the data generated by this method still lie within the original sample's coverage space and cannot extend local space samples to the global space. For problems with relatively clear mechanistic models, machine learning methods may not be the best choice.

Sample generation methods based on information diffusion criteria are more reliable than machine learning methods. These methods use a mathematical approach based on statistical theory and were first proposed by Huang (1997). The mega-trend-diffusion (MTD) method and its related variants, which evolved from this theory, are highly effective virtual sample generation strategies. Yu et al. (2019) proposed the MTD and Monte Carlo virtual sample generation method to address the small sample problem. Khamis et al. (2022) studied the optimization of MTD parameters, enhancing the quality of sample generation. Sivakumar et al. (2022) integrated k-nearest neighbors into MTD, optimizing its application to small sample scenarios. In this study, we used trend diffusion as a fundamental method to generate virtual samples for fault diagnosis in building air conditioning systems, aiming to enhance sample information completeness and address the dependency of data-driven fault diagnosis methods on high-quality labeled data.

1.2 Related work for fault detection and diagnosis

To analyze the significance and future trends of this study from a broader perspective, in this section we provide a comprehensive review of fault detection and diagnosis (FDD) methods in the field of building energy systems. Generally, these methods can be categorized into two main types: interpretable and non-interpretable. Interpretable FDD methods involve diagnostic processes that are understandable to humans, including techniques such as knowledge graphs (Lu et al., 2022), knowledge reasoning (Zhang et al., 2022), statistical methods (Chen et al., 2021), Bayesian networks (Wang et al., 2017, 2021; Li et al., 2022b), and decision trees (Yu et al., 2018). In contrast, non-interpretable

FDD methods involve mainly black-box approaches based on neural networks, with backbone models further divided into types such as convolutional neural network (CNN) (Sun et al., 2019), recurrent neural network (RNN) (Sun et al., 2020), and transformer (Sun et al., 2022).

Fifteen years ago, the mainstream FDD methods in building energy systems were interpretable, like Bayesian networks. However, these methods faced significant limitations in addressing complex building systems due to the difficulty of acquiring detailed expert knowledge. With the rapid development of deep learning, black-box diagnostic methods based on deep neural networks have quickly come to dominate the FDD field, effectively tackling FDD challenges in complex building systems through autonomous data learning. However, black-box methods have two major limitations: data dependency and lack of interpretability (Zhao et al., 2019; Chen et al., 2023). Data dependency restricts their application in data-scarce environments, while the lack of interpretability limits their use in areas with strict reliability requirements.

In recent years, more researchers have begun to explore the integration of both diagnostic approaches, aiming to achieve FDD through both knowledge and data. This approach not only meets reliability requirements but also adapts to the complexities of modern building systems. Knowledge and data-driven FDD methods encompass various techniques, with three receiving the most attention: physics-informed neural networks (PINNs) (Wei and Ooka, 2023), knowledge graphs (Li et al., 2022a), and knowledge-guided virtual samples (Sun and Yao, 2024; Sun et al., 2024). These methods have their own strengths and weaknesses and are applicable in different scenarios. This paper focuses on a knowledge-guided virtual sample generation method based on trend diffusion, as described in Section 1.1. However, the current state of research reveals that the core trend diffusion method is the use of a small number of samples to estimate the diffusion boundary, raising two critical issues: (1) What if there are no small samples? (2) What if the small samples are biased in their distribution? Addressing these two issues became the motivation for our research.

1.3 Motivation

Current virtual sample generation strategies are based on small sample conditions. Small sample sets, being subsets of the overall population, represent only partial observations, often leading to distribution bias (Lu and Li, 2013). This bias further affects the distribution of generated samples. Moreover, in some scenarios, small samples are difficult to obtain, necessitating sample generation under zero-sample conditions.

Based on the above analysis, in this paper we address two key questions: How can samples be generated under zero-sample conditions? How can the completeness of information in generated samples be ensured? To address these issues, a knowledge-guided virtual sample generation method is proposed to enhance information completeness and facilitate zero-sample fault diagnosis. Niyogi et al. (1998) mathematically proved that virtual samples constructed using domain knowledge can expand a training set's information similarly to real samples. However, the challenge lies in acquiring the necessary domain knowledge for sample generation.

With the release of ChatGPT 3.5 in 2022, AI development has entered the era of large models, with various general-purpose large models rapidly emerging, providing new avenues for obtaining domain knowledge. Scholars have rapidly advanced research into the application of LLMs within the building energy sector. Zhang et al. (2024a) leveraged generative pre-trained transformers (GPT) to develop an automated data mining framework aimed at enhancing building energy conservation, enabling the automated extraction of knowledge from large datasets. Lu et al. (2024) conducted a comprehensive evaluation of the proficiency of various LLMs in heating, ventilation, and air conditioning (HVAC)-related knowledge, revealing that GPT-4 was capable of passing the ASHRAE Certified HVAC Designer examination, with a performance surpassing that of about half of the human examinees. Jiang et al. (2024) introduced an LLM-based computing platform designed for automated building energy modeling, which uses LLMs for automatic knowledge description and the generation of intermediate data format (IDF) files. Forth and Borrmann (2024) applied LLMs to automatically fill in missing information in building information models, effectively showcasing the expertise of LLMs in the building energy field. These studies showed convincingly that LLMs have sufficient expertise to output professional knowledge at a level approaching that of human experts. The recent groundbreaking study by Zhang et al. (2025) used fine-tuned LLMs for HVAC fault diagnosis. Through an optimized fine-tuning strategy, the model is transformed into a domain-specific large model, significantly boosting diagnostic accuracy. This innovation can be deployed in the cloud for continuous monitoring. However, because LLMs require substantial computational resources, they are less suitable for local or embedded monitoring scenarios. In contrast, our knowledge-guided mega-trend-diffusion (KG-MTD) method leverages the broad foundational knowledge of LLMs and uses heuristic questioning to extract the domain knowledge needed for sample generation. Training a diagnostic model with these generated samples not only ensures ample domain knowledge but also keeps computational demands relatively low.

Research has identified notable differences between virtual samples and real samples, due mainly to the relatively low information entropy of the virtual samples, which results in poor generalization of the trained models. Sample information entropy measures the degree of perturbation present; increasing this entropy can enhance the training effectiveness of neural networks. Real samples often have higher information entropy due to the presence of real-world perturbations. While these perturbations may negatively influence measurements, they can be beneficial to some extent for neural network training (Bishop, 1995). Holmstrom and Koistinen (1992) proposed that introducing noise into training samples could improve neural network training effectiveness by increasing sample information entropy. Zhou et al. (2019) validated the mathematical theory behind noise addition at the International Conference on Machine Learning, further confirming its efficacy. Building on this theory, in this paper, we enhance the quality of virtual sample generation by adding Gaussian noise to the virtual samples and optimizing the addition parameters, thereby developing the proposed virtual sample generation method.

1.4 Contributions

The main contributions of this paper are as follows:

(1) An improved MTD method is proposed, leveraging domain knowledge acquired from LLMs to guide the design of diffusion function boundaries, thereby reducing dependence on small samples while ensuring the global diffusion function is unbiased.

(2) An optimized diffusion probability density function is developed, and Monte Carlo sampling is used to generate virtual samples, creating an information-complete virtual sample set.

(3) A noise-adding optimization strategy is introduced to boost the information entropy of the virtual sample set, enhancing the robustness and generalization of neural network models.

2 Methodology

In this section, we briefly describe the methods and principles of knowledge-guided virtual sample generation. It consists of the following four steps: domain knowledge acquisition, diffusion function design, virtual sample generation, and virtual sample enhancement. The implementation procedure of the proposed method is illustrated in Fig. 1. The details of the implementation of each step are described below.

2.1 Basic theory

This section briefly introduces the basic theories of the method proposed in this paper, including MTD and Monte Carlo sampling.

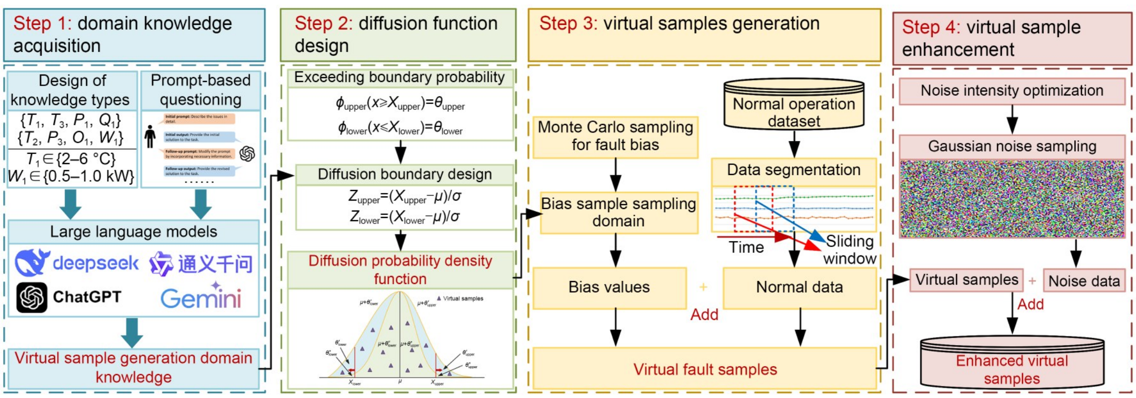

Fig. 1 Framework of the proposed methodology. All the variables will be explained in the main text

2.1.1 Mega-trend-diffusion

MTD is a technique used to estimate the acceptable range of sample attributes and diffuse data information (Yu et al., 2019). Li et al. (2007) first proposed MTD to address the small sample problem in neural network training. The fundamental concept of MTD originates from the information diffusion theory proposed by Huang and Moraga (2004), which was initially developed to address the issue of information completeness in small samples. The diffusion formula was shown as:

$$x_i = u_i \pm \sqrt{-2 \times h \times \ln(\varphi(x_i))}, \tag{1}$$

where x_i is the estimated sample, u_i is the collected sample, and h is the diffusion coefficient. $\varphi(x_i)$ is the value of the membership function of x_i.

The diffusion coefficient h is estimated based on the sample set to define the upper and lower limits of diffusion. Li et al. (2007) provided the specific method for estimating h.

Let h_{set}, the set diffusion coefficient, be

$$h_{\text{set}} = \frac{\hat{s}_x^2}{n}, \tag{2}$$

where $\hat{s}_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)$ is the sample set variance, n is the sample size, and u_{set} is the core of the sample set:

$$u_{\text{set}} = (x_{\min} + x_{\max})/2, \tag{3}$$

where $x_{\min}$ represents the minimum sample value, and $x_{\max}$ represents the maximum sample value.

Thus, the lower and upper limits of the diffused sample set B_{L} and B_{U} are:

$$B_{\text{L}} = u_{\text{set}} - \sqrt{-2 \times \frac{\hat{s}_x^2}{N_{\text{L}}} \times \ln(\varphi(B_{\text{L}}))}, \tag{4}$$

$$B_{\text{U}} = u_{\text{set}} + \sqrt{-2 \times \frac{\hat{s}_x^2}{N_{\text{U}}} \times \ln(\varphi(B_{\text{U}}))}, \tag{5}$$

where N_{L} and N_{U} represent the number of data smaller and greater than u_{set}, respectively.

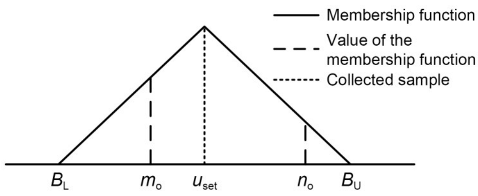

Fig. 2 Mega-trend-diffusion technique. m_o and n_o represent the values of the membership function during the diffusion process, and the probabilities of these values influence the diffusion outcome

Furthermore, to address the issue of asymmetric domain range expansion, the skewness coefficients are introduced into the diffusion process.

$$S_{\text{L}} = \frac{N_{\text{L}}}{N_{\text{L}} + N_{\text{U}}}, \tag{6}$$

$$S_{\text{U}} = \frac{N_{\text{U}}}{N_{\text{L}} + N_{\text{U}}}. \tag{7}$$

The diffusion functions are accordingly revised as:

$$B_{\text{L}} = u_{\text{set}} - S_{\text{L}} \times \sqrt{-2 \times \frac{\hat{s}_x^2}{N_{\text{L}}} \times \ln(\varphi(B_{\text{L}}))}, \tag{8}$$

$$B_{\text{U}} = u_{\text{set}} + S_{\text{U}} \times \sqrt{-2 \times \frac{\hat{s}_x^2}{N_{\text{U}}} \times \ln(\varphi(B_{\text{U}}))}. \tag{9}$$

Ultimately, MTD generates new data points in an asymmetric diffusion manner using membership functions and data samples, achieving the purpose of dataset augmentation.

2.1.2 Monte Carlo sampling

Monte Carlo sampling techniques provide solutions to large-scale stochastic sampling problems. Compared to deterministic algorithms like the Las Vegas algorithm, which consistently produce precise outcomes, the Monte Carlo algorithm introduces an element of random error into its results. However, this error can be mitigated by allocating more computational resources. In many machine learning tasks, precisely defining the specific problem is often impractical, rendering deterministic algorithms unfeasible. In such instances, the Monte Carlo algorithm proves to be a valuable estimation tool. The technique of sampling in space using the Monte Carlo algorithm is known as Monte Carlo sampling.

The Monte Carlo sampling method involves the following:

$$\Omega=\sum_{x}p(x)f(x)=\mathbb{E}_p[f(x)],\tag{10}$$

where Ω is real space, p is the probability distribution of the random variable x, f is the function defined over the sampling space, and $\mathbb{E}_p$ is a mapping function from f to Ω. We can approximate a quantity $\hat{\Omega}$ by taking v samples $x^{(1)}$, $x^{(2)}$, $\cdots$, $x^{(v)}$ from p, and obtain an empirical mean:

$$\hat{\Omega}=\frac{1}{v}\sum_{i=1}^{v}f(x^{(i)}).\tag{11}$$

According to the law of large numbers, if the samples $x^{(i)}$ are independent and identically distributed, their average will converge to the expected value as the number of samples increases:

$$\lim_{v\to\infty}\frac{1}{v}\sum_{i=1}^{v}f(x^{(i)})=\mathbb{E}_p[f(x)].\tag{12}$$

Thus, Monte Carlo sampling provides a cost-effective means to approximate complex functions over high-dimensional spaces.

2.2 Domain knowledge for virtual sample generation

2.2.1 Framework of requirements for domain knowledge

In thermal energy systems, such as building air conditioning systems, certain parameters typically deviate from normal operating conditions during faults. For example, refrigerant leakage can cause a decrease in condensing pressure and a reduction in subcooling temperature, which serve as crucial indicators for fault diagnosis. However, the range of parameter variations under normal operating conditions often exceeds those caused by faults due to the variable nature of thermal systems. This discrepancy can obscure fault deviation characteristics under dynamic conditions, complicating the direct application of fault deviation knowledge for diagnosis.

Nevertheless, recognizing these fault deviation patterns can be advantageous for generating virtual samples. By integrating deviation knowledge into the sample generation process, virtual samples that mimic fault deviation patterns can be produced, effectively simulating actual fault conditions. This method of knowledge-guided generation not only enhances overall credibility but also reduces reliance on actual data samples significantly.

Based on the fault characteristics in thermal systems, we have identified three categories of knowledge that are easily accessible and significantly improve the sample generation process. The details are provided in Table 1.

The aforementioned three categories of domain knowledge succinctly describe the variation patterns of faults in thermal systems. However, due to their lack of detail and constraints, they are only partially applicable in certain scenarios. Thus, it is essential to apply this knowledge within a fuzzy theory framework, using fuzzy probability analysis to estimate specific result classifications under conditions that are only partially met. Addressing this challenge through the use of virtual samples to drive neural networks for fault diagnosis was one of the main objectives of this study.

2.2.2 Approaches to acquiring domain knowledge

Building thermal systems, such as air conditioning, involve complex heat and mass transfer processes (Xu YJ et al., 2023; Li L et al., 2024a; Wu et al., 2024), making it challenging to accurately quantify dynamic changes in system parameters during faults. This limitation hampers the effectiveness of knowledge-driven

Table 1 Knowledge for virtual sample generation in building air conditioning systems

Knowledge type	Specific explanation	Example list
Change parameter set (K1)	A set of fault-related indicative parameters, representing the minimal parameter set accurately describing the fault. Different faults correspond to different parameter sets	$\{T_1, T_3, P_1, Q_1\}$; $\{T_2, P_3, O_1, W_1\}$
Deviation range (K2)	The specific possible variation range for each parameter in the set, corresponding mainly to different operating conditions, environments, and control requirements	$T_1\in\{2–6\ °C\}$, $W_1\in\{0.5–1.0\ kW\}$
Deviation mode (K3)	The mode value adjusted for deviation, providing a central tendency measure that takes into account any inherent bias in the data distribution	$T_1=5\ °C$ $W_1=0.82\ kW$

T: temperature; P: pressure; Q: flow; O: openness; W: power fault diagnosis methods. However, the method proposed in this paper bypasses the need for precise domain knowledge by using broad parameter change ranges to guide trend diffusion, generating virtual samples that statistically resemble real ones. Machine learning then adapts to these samples to perform tasks like fault diagnosis in a manner akin to fuzzy control. While such qualitative domain knowledge is readily accessible to experts, HVAC specialists are not always available in practical settings. Building operations teams, often composed of software engineers and managers, may consult HVAC experts occasionally, thereby incurring costs. LLMs offer a solution by providing essential domain knowledge without the need for constant expert consultation, saving both time and money.

Lu et al. (2024) conducted detailed tests on 12 mainstream LLMs in the HVAC field, analyzing their professional capabilities from the perspectives of recall, analysis, and application. The results showed that GPT-4 passed the ASHRAE Certified HVAC Designer Examination, scoring higher than half of the human test-takers. This study demonstrated that LLMs have knowledge levels comparable to human professionals in the HVAC field, effectively meeting the knowledge acquisition needs of the method proposed here. In this study, the GPT-4o model, an updated and more specialized version of GPT-4, was used for knowledge acquisition. Using GPT-4o to acquire domain knowledge and generate virtual samples is a key innovative approach of this research.

To obtain more accurate domain knowledge, it is necessary to describe relevant information about the air conditioning system to LLMs. Typically, this information needs to include the system structure, operating environment, control objectives, and refrigerant type. The more detailed the information, the more beneficial it is to the acquisition of accurate domain knowledge. Since LLMs are large models that process natural language, all information can be described using natural language, greatly simplifying the complexity of the knowledge acquisition process. The domain knowledge required by our proposed method is a form of universally applicable simplified knowledge, and our testing showed that the description content required by LLMs is actually very brief. Below is an example:

A chiller with a rated cooling capacity of 316 kW utilizes R134A as the refrigerant, with both the condenser and evaporator configured as water-to-water shell-and-tube heat exchangers. Under summer cooling conditions, with outdoor temperatures ranging from 35–40 °C, the chilled water outlet temperature is set at 7 °C, and the cooling water return temperature is set at 30 °C.

The above description already provides the most basic information required for knowledge generation. Of course, more detailed information is advantageous for obtaining more accurate knowledge.

To enhance domain knowledge acquisition, it is essential to design an effective questioning pattern that enables the LLM to accurately understand the problem. Here, we have devised four tricks to obtain more reliable responses.

Trick 1 (role setting): instruct the LLM about your role and current task. For example, you could say, “You are a senior HVAC engineer, and you are working on creating a universally applicable HVAC fault operation and maintenance manual.” This framing helps focus the model’s responses within the relevant context.

Trick 2 (task decomposition): break down a complex task into simpler, related questions. After defining the scenario, ask specific questions about each fault condition, focusing only on a few related parameters at a time. For example, when cooling water flow is reduced, you could inquire about the changes in condensing temperature and compressor discharge superheat.

Trick 3 (adversarial responses across multiple models): ask the same question to different models, then present the different responses to the LLM and ask it to assess which is the most accurate, encouraging the model to provide a better, more refined answer.

Trick 4 (repeated questioning): modify certain conditions and ask the same question repeatedly. For example, increase the ambient temperature by 1 °C each time and observe whether the model’s answers remain consistent with physical principles. You can also input the model’s response into another LLM for further evaluation.

By integrating the above prompting strategies, the reliability and accuracy of LLMs’ responses can be improved.

2.3 Proposed virtual sample generation method

2.3.1 Knowledge-guided mega-trend-diffusion (KG-MTD)

The primary feature of MTD is its use of real data samples to establish acceptable diffusion boundaries, within which new samples are generated. However, the value of MTD is limited in scenarios where obtaining real samples is challenging. To address this limitation, we propose KG-MTD. KG-MTD uses prior knowledge to establish diffusion boundaries, thereby eliminating the need for real samples and enabling the generation of virtual samples driven solely by knowledge.

Fig. 3 is a technical schematic of KG-MTD. The deviations of key parameters and their respective upper and lower boundaries are derived from LLMs and used as prior knowledge for MTD. Given that the deviation values of parameters in building thermal systems under fault conditions vary with factors such as the operating conditions, environment, and control requirement, and considering the inherent fuzziness of LLM knowledge, the diffusion process should not be strictly confined within the LLM-established boundaries. Consequently, we adopted the normal distribution as the diffusion function, replacing the triangular distribution used in MTD. This approach aligns with the characteristics of thermal system faults and ensures a certain probability of sampling beyond the boundaries.

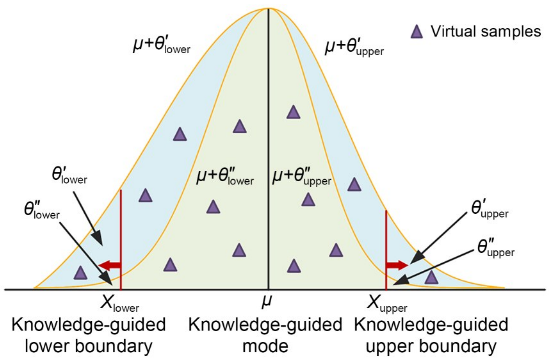

Fig. 3 Knowledge-guided mega-trend-diffusion technique. All variables will be explained in the maintext

Due to the variable deviation ranges and modes of key parameters, it is necessary to set the diffusion function individually for each parameter. Additionally, the asymmetric nature of upward and downward diffusion necessitates an asymmetric diffusion manner, with distinct diffusion functions designed for each direction.

The design of the diffusion function involves mainly establishing the optimal standard deviation σ of the normal distribution. This design is based on the upper and lower boundaries provided by the LLMs. Our design approach involves artificially setting a standard deviation beyond the normal distribution so that its probability density function satisfies a set probability outside the boundary range. Specifically, this is described as follows:

Let ϕ_{upper} be the probability density function for upward diffusion and ϕ_{lower} be the probability density function for downward diffusion. The sampling probability of the upward diffusion exceeding the upper boundary X_{upper} is θ_{upper}, and the sampling probability of the downward diffusion falling below the lower boundary X_{lower} is θ_{lower}. Therefore, we have:

$$\phi_{\text{upper}}(x \geqslant X_{\text{upper}}) = \theta_{\text{upper}}, \tag{13}$$

$$\phi_{\text{lower}}(x \leqslant X_{\text{lower}}) = \theta_{\text{lower}}. \tag{14}$$

The values of θ_{upper} and θ_{lower} corresponding to the standard normal distributions Z_{upper} and Z_{lower} can be obtained from the standard normal distribution table. Using these values, Z can be converted into X in the diffusion function as follows:

$$Z_{\text{upper}} = \frac{X_{\text{upper}} - \mu}{\sigma}, \tag{15}$$

$$Z_{\text{lower}} = \frac{X_{\text{lower}} - \mu}{\sigma}, \tag{16}$$

where X is the boundary value and μ is the mode given by the LLMs. Therefore, the standard deviation of the normal diffusion function can be calculated by:

$$\sigma = \frac{X_{\text{upper}} - \mu}{Z_{\text{upper}}}, \tag{17}$$

$$\sigma = \frac{X_{\text{lower}} - \mu}{Z_{\text{lower}}}. \tag{18}$$

Consequently, the diffusion functions of KG-MTD, $X \sim N(\mu, \sigma^2_{\text{upper}})$ and $X \sim N(\mu, \sigma^2_{\text{lower}})$, can be derived.

Additionally, attention must be paid to the issue of the diffusion process crossing the zero point. For instance, if the upper and lower boundaries of a temperature deviation are (1, 3) and the mode is 2, with $\sigma_{\text{lower}} = 1$ after design, the diffusion function has a 2.28% probability of taking values less than 0. This would alter the deviation pattern of the parameter, which is undesirable. Therefore, when the issue of crossing the zero point arises, sampling should be performed as follows:

$$x = \begin{cases} \max(X_{\text{lower}}, 0), & \text{when downward diffusion,} \\ \min(X_{\text{upper}}, 0), & \text{when upward diffusion.} \end{cases} \tag{19}$$

2.3.2 Noise addition strategy for enhancing sample information entropy

Monte Carlo sampling within the established diffusion function space can generate virtual samples. However, neural networks trained with these virtual samples show poor generalization performance. We attribute this phenomenon to the low information entropy of the virtual samples.

Information entropy is a fundamental concept in information theory that quantifies uncertainty or the amount of information. Introduced by Claude Shannon in 1948, it is also known as Shannon entropy (Gappmair, 1999). The definition of information entropy is as follows:

$$H(X)=-\sum_{i} p(x_i)\log p(x_i), \tag{20}$$

where $H(X)$ is the entropy of the random variable X, and $p(x_i)$ is the probability of X taking the value x_i.

Several studies have demonstrated that the information entropy of data samples used to train neural networks significantly impacts network performance (Frank and Frank, 2020). Specifically, increasing the information entropy of training data enhances neural network performance. Holmstrom and Koistinen (1992) discovered that adding noise to the training dataset improved neural network training performance (Holmstrom and Koistinen, 1992). Zhou et al. (2019) explained the fundamental theory behind noise improving the effectiveness of neural network training. Practical experience has shown that introducing noise into training samples significantly enhances neural network training performance by increasing the information entropy of the training samples.

Based on this theory, we propose adding noise to the virtual samples generated by KG-MTD to improve neural network training effectiveness. Gaussian noise was selected because of its common use (Zhou et al., 2019), with noise intensity being the key optimization factor. Typically, the noise intensity should be related to the amplitude of the original signal, which is a topic that warrants further discussion.

2.3.3 Procedure for the proposed method

We summarize the procedure for implementing the proposed method in the following steps:

Step 1: Select an appropriate LLM to acquire the necessary knowledge for sample generation, including the change parameter set, deviation range, and deviation mode.

Step 2: Construct the trend diffusion function using the acquired knowledge and establish parameters such as the standard deviation for the bidirectional diffusion functions.

Step 3: Apply the Monte Carlo method to sample within the diffusion space and obtain virtual samples of parameter biases.

Step 4: Superimpose the virtual bias samples onto the corresponding parameters of real healthy samples to simulate fault biases, thereby generating virtual fault samples.

Step 5: Add Gaussian noise to the virtual fault samples to enhance their information entropy.

3 Case studies

3.1 Description of the platform in the case studies

3.1.1 Platform of air conditioning system

Here, we present experimental research based on the ASHRAE RP-1043 dataset, a widely used resource in chiller and AHU FDD studies. The experiments were conducted on a 90-t (316-kW) R134a centrifugal chiller in a room maintained at 70 °F (21.1 °C). A system schematic is shown in Fig. 4. Nine typical faults were artificially induced, each applied at four severity levels ranging from 10% to 40%. The experiments ran under 27 different operating conditions, lasting about 14 h per fault level. Operational data for 64 variables (e.g., temperature, pressure, flow rate, power) were collected at 10-s and 2-min intervals. Each test began with a

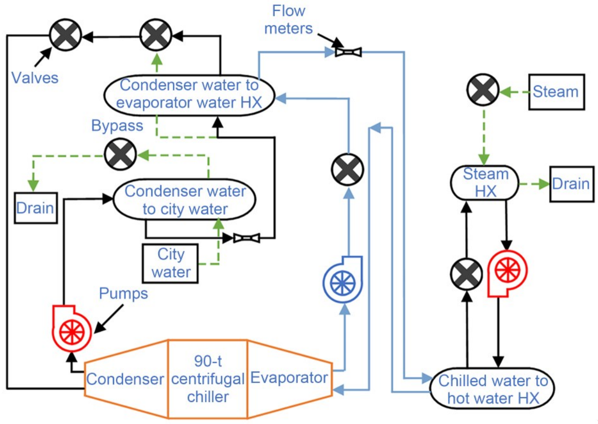

Fig. 4 Schematic diagram of chiller structure

30-min stabilization period, followed by 15–25 min of steady-state data collection.

We selected six types of faults and one type of healthy data from the project. The six fault types were: condenser fouling (cf), reduced condenser water flow (fwc), reduced evaporator water flow (fwe), non-condensables in refrigerant (nc), refrigerant leak (rl), and refrigerant overcharge (ro). The simulated conditions for different severity levels of each fault are shown in Table 2. In these simulations, "blocked tubes" refer to the number of blocked heat exchanger tubes, "gpm" stands for gallons (1 gallon=3.78541 L) per min, indicating the water flow rate, and "pound" (1 pound≈0.453592 kg) measures the weight of refrigerant. This dataset required normalization preprocessing in subsequent tests. The specific algorithm is as follows:

$$x'=\frac{x-\mu}{3\sigma^2}, \tag{21}$$

where x represents the collected sample, x' represents the normalized sample, μ represents the mean, and σ^2 represents the variance.

Comstock et al. (2001) investigated the sensitivity and deviation characteristics of the features of this dataset, identifying seven key features for fault representation: temperature difference between the inlet and outlet of the evaporator ($T_{\text{EI-EO}}$), temperature difference between the inlet and outlet of the condenser ($T_{\text{CO-CI}}$), evaporator pressure (P_{E}), condenser pressure (P_{C}), subcooling degree (T_{sub}), suction superheat (T_{suc}), and discharge superheat (T_{dis}). ASHRAE RP-1043, a seminal open-source dataset for building thermal system faults, serves as a highly representative benchmark. We used this dataset as a case study to assess the effectiveness of the proposed methodologies.

3.1.2 Platform for algorithm operating

The deep learning algorithm was implemented using TensorFlow 2.6.0 and Python 3.9.7 within the PyCharm 2022 development environment. It uses CUDA 11.2 and cuDNN 8.1 libraries for GPU acceleration. The algorithm is executed on a graphics server equipped with an NVIDIA GeForce RTX 3080Ti GPU, Intel i9-11900K CPU, 64 GB of RAM, and a 64-bit Windows 10 operating system.

3.2 Experimental design

We designed five different sets of experiments to achieve optimization analysis and comparative validation of the proposed method.

3.2.1 Experiment 1: comparative analysis of machine learning classifiers

Objective: to compare the performance of five machine learning classifiers—CNN, ANN, support vector machine (SVM), Random Forest, and XGBoost—for HVAC fault diagnosis, and identify the optimal baseline model for further research.

These five machine learning algorithms are commonly used in the HVAC domain (Zhao et al., 2019). Comparing their performance in this case study will benefit subsequent research. All algorithms use a feature

Table 2 Simulated conditions for faults of variable severity

Fault type	Simulated condition			
	Level 1	Level 2	Level 3	Level 4
cf	12% reduction in tubes (20 blocked tubes)	20% reduction in tubes (33 blocked tubes)	30% reduction in tubes (49 blocked tubes)	45% reduction in tubes (74 blocked tubes)
fwc	10% reduction in flow (243 gpm)	20% reduction in flow (216 gpm)	30% reduction in flow (189 gpm)	40% reduction in flow (162 gpm)
fwe	10% reduction in flow (194 gpm)	20% reduction in flow (173 gpm)	30% reduction in flow (151 gpm)	40% reduction in flow (130 gpm)
nc	1% by volume nitrogen (0.10 pound)	2% by volume nitrogen (0.16 pound)	3% by volume nitrogen (0.22 pound)	5% by volume nitrogen (0.54 pound)
rl	10% reduction in charge (270 pound)	20% reduction in charge (240 pound)	30% reduction in charge (210 pound)	40% reduction in charge (180 pound)
ro	10% increase in charge (330 pound)	20% increase in charge (360 pound)	30% increase in charge (390 pound)	40% increase in charge (420 pound)

parameter matrix consisting of 25 time steps as input. Since conventional classifiers cannot directly process matrix data, non-CNN models use a flattening algorithm to convert the two-dimensional matrix into a one-dimensional feature vector. Notably, the classic SVM was optimized with a linear kernel function due to computational time constraints caused by high-dimensional data.

The training dataset was generated via numerical simulation to create a virtual sample set: 1000 samples were selected for each of the seven fault modes (a total of 7000 samples), while the test dataset consisted of actual operational data (500 samples per class, totaling 3500 samples). The virtual sample generation strategy was based on empirical methods, where the upper and lower bounds of the parameters were set to $3\sigma_{\text{upper}}$ and $1.5\sigma_{\text{lower}}$, and 5% Gaussian noise was added to simulate real operational disturbances. This strategy was intended to ensure a fair performance comparison among different classifiers.

3.2.2 Experiment 2: optimization analysis of generation strategy

Objective: to determine the optimal generation strategy by optimizing the intensity of noise and the distribution of the diffusion function.

The intensity of noise addition and the distribution of the diffusion function are two key hyperparameters that affect the quality of generated data. This experiment aimed to obtain the optimal values for these hyperparameters. Additionally, to ensure the generalizability of these hyperparameters across different scenarios, it was necessary to apply certain processing methods to decouple them from the specific context. The evaluation process involved training a CNN-based diagnostic model using virtual samples. For each fault type, 1000 sets of virtual samples were selected, resulting in a total of 7000 virtual samples. The model was subsequently tested on a real dataset, with the test set comprising 500 samples per fault type, totaling 3500 samples.

The main aim of adding noise is to enhance the information entropy of the virtual samples. However, excessive noise can distort the samples, while insufficient noise fails to sufficiently increase information entropy. Thus, optimizing the intensity of noise addition is crucial.

For a given Gaussian noise $N(\mu, \sigma^2)$, its distribution is determined by the mean μ and variance σ^2. For the task of adding noise to virtual samples, the noise should not have a fixed bias, meaning it needs to satisfy $\mu=0$, hence the parameter to be optimized is the variance σ^2. The composition of the virtual sample can be expressed as:

$$D_{\text{virtual}} = D_{\text{normal}} + D_{\text{deviation}}, \tag{22}$$

where D represents the sample data variable, D_{virtual} is the virtual sample, D_{normal} is the normal sample, and $D_{\text{deviation}}$ is the generated deviated sample.

$D_{\text{deviation}}$ was sampled from the diffusion function space, with the mean of its diffusion function being the mode obtained from the LLMs. Using the mode value M as the baseline, we defined the noise intensity as follows: when $\sigma^2=M/10$, the noise intensity was 10%; when $\sigma^2=M$, the noise intensity was 100%. This method of defining noise intensity based on the mode value effectively achieves the goal of decoupling from the specific scenario. We validated the quality of virtual samples at noise intensities of 0%, 2%, 5%, 10%, 20%, 40%, and 100%. To ensure the single-variable principle in the experiment, we used a unified diffusion function distribution with both upper and lower standard deviations equal to $X/2$ (the boundary value provided by the LLMs).

The diffusion function serves as the sampling space for generating virtual samples, and its distribution is of significant importance for the informational completeness of the virtual samples. In this method, the diffusion function follows a normal distribution, with its mean being the deviated mode obtained from LLMs. The standard deviation of the normal distribution is the parameter that needs to be optimized.

In the known domain knowledge, there are upper and lower deviation boundaries X_{upper} and X_{lower} for each parameter. However, in actual sampling, we do not wish to sample strictly within this range. Therefore, by setting an appropriate standard deviation σ, we aimed to obtain virtual samples that can exceed the range with a certain probability. The setting of σ should be based on X, and the distribution of the upper and lower deviations can be asymmetric. We represent the standard deviation of the diffusion function distribution in the form of $X=a\sigma$, where a is the coefficient for calculating the standard deviation (CCSD) that needs to be optimized. For example, $X=3\sigma$ means the boundary value provided by the LLMs equals three times the standard deviation of the distribution, implying a 0.3% probability of exceeding the boundary.

To optimize the diffusion function, we conducted 20 trials with different CCSDs, labeled S1–S20. The upward CCSD and the downward diffusion CCSD were as follows (b_{upper}/b_{lower}): S1-1/3, S2-1/2.5, S3-1/2, S4-1/1.5, S5-1/1, S6-1.5/3, S7-1.5/2.5, S8-1.5/2, S9-1.5/1.5, S10-1.5/1, S11-2/3, S12-2/2.5, S13-2/2, S14-2/1.5, S15-2/1, S16-3/3, S17-3/2.5, S18-3/2, S19-3/1.5, and S20-3/1.

3.2.3 Experiment 3: diagnostic analysis based on optimal strategy

Objective: to evaluate the quality of virtual samples using assessment metrics on benchmark datasets.

The optimal virtual sample generation strategy was identified through Experiment 2. However, the validation in that experiment lacked detail and did not demonstrate the specific diagnostic performance for different faults. In this experiment, we used a benchmark dataset to conduct detailed validation for various faults and their severities, obtaining the accuracy, recall, and F1 score metrics for each type of fault.

The model training process in this study was consistent with that of Experiment 3. The diagnostic model was trained using 7000 sets of virtual samples, and the evaluation metrics were calculated on 3500 sets of real data.

3.2.4 Experiment 4: comparison with supervised learning on real samples

Objective: to validate the superiority of this method by comparing it with models trained on real samples.

The main advantage of the proposed method is that it does not require real fault data and enhances the completeness of sample information, thereby addressing data dependency and insufficient generalization issues in deep learning. However, the difference in performance between models trained with real data and those trained with virtual data remained a concern. To address this, six sets of comparative tests were designed, each containing real data of variable severity, to verify the differences in performance between diagnostic models trained with real samples and those trained with virtual samples under conditions of a small sample size and incomplete information. The specific experimental design was shown in Table 3.

Table 3 Comparative design of Experiment 3

Group	Data source of training set	Sample size per fault type	Training set sample size
1	Virtual dataset	1000	7000
2	Level 1 dataset	1000	7000
3	Level 2 dataset	1000	7000
4	Level 3 dataset	1000	7000
5	Level 4 dataset	1000	7000
6	Levels 1–4 dataset	250	7000

3.2.5 Experiment 5: impact of LLM output knowledge accuracy on diagnostic performance

Objective: to quantitatively evaluate the impact on diagnostic outcomes when the knowledge output by LLMs exhibits numerical or directional biases.

Since the knowledge generated by LLMs is not always entirely accurate, it is essential to investigate the extent to which deviations in this output affect the final diagnostic results. The knowledge obtained from LLMs consists of parameter deviations caused by system faults, where the magnitude of these deviations is influenced by factors such as fault severity, ambient temperature, and operating conditions. Consequently, the knowledge provided by LLMs inherently represents a reference range rather than an exact value. In this study, we quantified the impact of fluctuations in the acquired knowledge on diagnostic performance from two perspectives.

First, we assumed that the directional deviations of each parameter provided by LLMs were correct, but the specific numerical values contained some errors. Building on the deviation knowledge output by LLMs, we randomly scaled the deviation magnitudes within a range of 50% to 200% to simulate potential knowledge inaccuracies from LLMs. Using these newly generated knowledge vectors, we constructed virtual samples to train a diagnostic model and evaluate its performance. The training and evaluation strategies for the diagnostic model remained consistent with those used in previous experiments. Each training and evaluation cycle was carried out seven times, with the average value taken as the performance metric for that cycle. The experiment comprised 50 simulation runs, generating 50 distinct knowledge vectors and yielding 50 corresponding accuracy rates.

Next, we explored the effects when the knowledge output by LLMs contained errors in the direction of parameter deviations. For this analysis, we selected a subset of parameters exhibiting deviations and reversed their directional bias to create new knowledge vectors with intentional errors. We examined four levels of error severity, defined by the proportion of parameters with reversed directions: 10%, 12%, 15%, and 25%. Virtual samples were then generated based on these vectors, and the diagnostic accuracy was subsequently assessed.

3.2.6 Evaluation metrics

The purpose of generating virtual samples is to enhance the performance of data-driven fault diagnosis models. We evaluated the quality of these virtual samples using direct diagnostic performance metrics. Specifically, we trained data-driven models with virtual samples and tested them on standard datasets. To comprehensively assess the performance of the fault diagnosis models, we used various metrics such as accuracy, recall, precision, and F1 score. The specific definitions of these metrics are as follows.

First, we defined four metrics used to evaluate classification models: true positive (TP), correctly predicted positive cases; false negative (FN), actual positives incorrectly predicted as negative; false positive (FP), actual negatives incorrectly predicted as positive; true negative (TN), correctly predicted negative cases.

Further, we used the above four metrics to define accuracy, recall, precision, and F1 score:

Accuracy (correct rate): measures the proportion of correctly predicted instances (both positives and negatives) out of all instances.

$$M_{\text{Accuracy}}=\frac{M_{\text{TP}}+M_{\text{TN}}}{M_{\text{TP}}+M_{\text{TN}}+M_{\text{FP}}+M_{\text{FN}}}. \tag{23}$$

Recall rate: measures the proportion of actual positive instances correctly identified by the model.

$$M_{\text{Recall}}=\frac{M_{\text{TP}}}{M_{\text{TP}}+M_{\text{FN}}}. \tag{24}$$

Precision: measures the proportion of positive predictions that are actually correct.

$$M_{\text{Precision}}=\frac{M_{\text{TP}}}{M_{\text{TP}}+M_{\text{FP}}}. \tag{25}$$

F1 score: the harmonic mean of precision and recall, providing a balance between recall and precision.

$$M_{\text{F1}}=\frac{2\cdot(M_{\text{Precision}}\times M_{\text{Recall}})}{M_{\text{Precision}}+M_{\text{Recall}}}. \tag{26}$$

The classification model uses a convolutional neural network as the backbone. By inputting the multi-parameter time series data of the chiller into the model in a matrix form and training it under supervision, we obtained an evaluation model that indirectly assessed the quality of the virtual samples.

4 Results and discussion

4.1 Acquiring domain knowledge using LLMs

Although previous studies have validated the expertise of LLMs in the HVAC field (Lu et al., 2025), it was necessary to assess whether LLMs can accurately answer the specific questions involved in this study. In particular, we wanted to know the extent of deviations in various thermodynamic parameters when faults occur. Typically, the degree of parameter deviation is related not only to the severity of the fault but also to factors such as system structure, refrigerant type, and operating conditions.

Firstly, we designed the question shown in Fig. 5 to analyze whether LLMs provide correct responses. For model selection, we used the two most powerful models currently available: ChatGPT o3-mini-high and DeepSeek R1. As shown in Fig. 5, when simple prompts were provided, the LLMs were able to offer suggested answers in a manner consistent with common sense. These responses covered the typical operating range of the system under normal conditions, the degree of deviation after a fault occurs, and the theoretical basis behind the answers. This indicates that the responses of the LLMs adhered to fundamental physical principles, thus demonstrating a degree of reliability. Although there were some differences in the prediction of specific numerical changes, the overall trend of parameter changes predicted by the LLMs was entirely correct.

Furthermore, to verify their understanding of the details of refrigeration principles, we altered the refrigerant type and heat exchanger type in the system and designed the question shown in Fig. 6. The results showed that the LLMs had relevant knowledge regarding the influence of different refrigerants on system conditions and adjusted the parameter changes accordingly, further confirming the reliability of their responses.

Prompt

You are a senior HVAC engineer, and you are working on creating a universally applicable HVAC fault operation and maintenance manual.

Consider the following scenario: a chiller with a rated cooling capacity of 316 kW uses R134A as the refrigerant, and both the condenser and evaporator are configured as water-to-water shell-and-tube heat exchangers. Under summer cooling conditions, with outdoor temperatures ranging from 25 to 35 °C, the chilled water outlet temperature ranges from 5 to 10 °C, and the cooling water return temperature ranges from 20 to 30 °C.

When the chiller experiences the following fault: non-condensable gases in refrigerant (1%–5% by volume nitrogen). Please analyze how the following two parameters will change: condenser pressure; discharge superheat.

Please note that we need to list the deviation range and average deviation of the two parameters relative to normal operating conditions when the fault occurs. For example, for fault *A*, parameter *X* will increase, with an increase range of approximately 3–5 °C and an average deviation of 3.8 °C. Please provide your reasoning.

Output of GPT o3-mini-high

1. Condenser pressure:

- **Change**: The presence of non-condensable gases reduce the effective condensation area in the condenser, leading to an increase in condenser pressure.
- **Deviation range**: The condenser pressure typically increases by about 3–5 bar compared to normal operating conditions.
- **Average deviation**: Approximately 4 bar, depending on the volume of non-condensable gas.

2. Discharge superheat:

- **Change**: Non-condensable gases affect condensation efficiency, causing an increase in discharge temperature, which in turn raises discharge superheat.
- **Deviation range**: Discharge superheat typically increases by about 5–10 °C.
- **Average deviation**: Approximately 7 °C, reflecting the impact of reduced condenser efficiency.

Reasoning:

Non-condensable gases hinder the heat transfer process in the condenser, leading to higher condenser pressure and more refrigerant remaining superheated, which results in an increase in discharge superheat.

Output of DeepSeek R1

1. Condenser pressure:

- **Change:** Non-condensable gases occupy space in the condenser, reducing effective heat transfer and increasing total system pressure.
- **Deviation range:** Condenser pressure increases by approximately 1.5–3.0 bar compared to normal conditions.
- **Average deviation:** About 2 bar, proportional to the non-condensable volume (1%–5% nitrogen).

2. Discharge superheat:

- **Change:** Elevated condenser pressure raises refrigerant saturation temperature. However, the compressor discharge temperature increases further due to reduced condenser efficiency, amplifying superheat.
- **Deviation range:** Discharge superheat increases by 8–15 °C.
- **Average deviation:** About 11 °C, reflecting impaired heat rejection.

Reasoning:

Non-condensable gases (e.g., nitrogen) increase the total pressure in the condenser by occupying volume without condensing. This reduces the effective partial pressure of the refrigerant, forcing the compressor to work harder to achieve the same cooling effect. The higher total pressure raises the refrigerant's saturation temperature, but the actual condensing efficiency drops, leading to a disproportionate rise in compressor discharge temperature.

Fig. 5 Verification of LLM capabilities in chiller fault diagnosis

Finally, to obtain more accurate values for parameter changes, we used prompt engineering techniques by synthesizing the answers from both LLMs, allowing the LLMs to independently assess and integrate the information in a more guided manner to derive the final answer. As shown in Fig. 7, after a comprehensive analysis, the LLMs were able to understand the strengths and weaknesses of the two answers and combine them into a more optimal solution. In conclusion, the LLMs are capable of understanding professional issues in the refrigeration field. Although their responses may not have been entirely accurate, the differences were typically not significant and the responses were sufficient to support the methods proposed in this paper. The relevant fault types and key parameters for the required domain knowledge are outlined below.

The relevant fault types are: condenser fouling (cf); reduced cooling water flow (fwc); reduced chilled water flow (fwe); presence of non-condensable gas (nc); refrigerant leakage (rl); refrigerant overcharge (ro).

The key parameters are: evaporator inlet-outlet temperature difference ($T_{\text{EI-EO}}$); condenser inlet-outlet temperature difference ($T_{\text{CO-CI}}$); evaporator pressure (P_{E}); condenser pressure (P_{C}); subcooling degree (T_{sub}); suction superheat (T_{suc}); discharge superheat (T_{dis}).

Fault severity descriptions are: cf: 10%–40% of heat exchanger tubes blocked; fwc: water flow reduced by 10%–40%; fwe: water flow reduced by 10%–40%; nc: presence of 1%–4% non-condensable gas; rl: refrigerant leakage of 10%–40%; ro: refrigerant overcharge of 10%–40%.

Finally, we used practical techniques, such as repeated questioning, to obtain the necessary domain

Prompt

Please continue answering from the previous question. If the refrigerant of the chiller is changed to R290, and both the evaporator and condenser are replaced with air-cooled heat exchangers, please correct the above answer.

Output of GPT o3-mini-high

1. Condenser pressure:

- **Change**: With R290 as the refrigerant and air-cooled heat exchangers, the presence of non-condensable gases will still hinder the condensation process, but the overall effect on condenser pressure may be slightly more pronounced due to the lower heat transfer efficiency of air-cooled systems compared to water-cooled systems.
- **Deviation range**: The condenser pressure will likely increase by about 4–6 bar compared to normal operating conditions.
- **Average deviation**: Approximately 5 bar, considering the combination of non-condensable gases and the air-cooled heat exchangers' lower efficiency.

2. Discharge superheat:

- **Change**: With air-cooled heat exchangers, the reduced heat transfer capability further exacerbates the effect of non-condensable gases, causing a more significant increase in discharge superheat.
- **Deviation range**: The discharge superheat will typically increase by 7–12 °C.
- **Average deviation**: Approximately 9 °C, reflecting the combined impact of non-condensable gases and the air-cooled system's reduced efficiency.

Reasoning:

Non-condensable gases still prevent proper condensation, but with R290 and air-cooled heat exchangers, the heat rejection capacity is lower, leading to more significant pressure and temperature increases. The air-cooled system struggles more to reject heat, causing higher condenser pressures and discharge superheats.

Output of DeepSeek R1

1. Condenser pressure:

- **Change:** Non-condensable gases (e.g., nitrogen) occupy space in the air-cooled condenser, reducing effective heat transfer. Air-cooled systems inherently operate at higher condensing pressures due to lower heat rejection efficiency compared to water-cooled systems. The lower operating pressure of R290 (vs. R134A) exacerbates the relative pressure rise caused by non-condensable gases.
- **Deviation range:** Condenser pressure increases by 3–6 bar compared to normal conditions.
- **Average deviation:** About 4.5 bar, influenced by air-cooled inefficiency and R290's lower baseline pressure.

2. Discharge superheat:

- **Change:** Air-cooled condensers struggle to reject heat, raising the refrigerant's saturation temperature. Non-condensable gases further reduce heat transfer, forcing the compressor to work harder. While R290 has better thermal properties (e.g., higher latent heat), the combined effect of air-cooled limitations and non-condensable gases amplifies discharge superheat.
- **Deviation range:** Discharge superheat increases by 15–25 °C.
- **Average deviation:** About 20 °C, reflecting compounded inefficiencies in heat rejection.

Reasoning:

- R290 properties: Propane (R290) operates at lower absolute pressures than R134A but is more sensitive to pressure changes due to its flammable nature and lower critical temperature. Air-cooled systems inherently run at higher condensing temperatures (e.g., 45–55 °C vs. 35–40 °C for water-cooled), reducing the margin for pressure increases.
- Non-condensable gases impact:
  - ○ Condenser pressure: Air-cooled condensers have poorer heat transfer efficiency, so non-condensable gases (1%–5% nitrogen) cause a larger pressure rise. The deviation range (3–6 bar) accounts for both the baseline pressure elevation in air-cooled systems and R290's sensitivity.
  - ○ Discharge superheat: Reduced condenser efficiency forces the compressor to operate at higher compression ratios, significantly elevating discharge temperatures. R290's lower specific heat capacity compared to R134A exacerbates superheat under fault conditions.

Fig. 6 Verification of LLM adaptability to system parameter changes

knowledge from LLMs, as summarized in Table 4. This knowledge aligns well with the parameter deviation patterns reported in related studies (Sun et al., 2022), indicating its reliability. This domain knowledge will be used in subsequent case studies to generate information-complete virtual samples.

4.2 Comparative analysis of various classification models

Fig. 8 illustrates the diagnostic performance of five different classification models on the evaluation dataset. Under identical training conditions, CNN delivered the best results in both correct rate and F1 score, achieving an overall diagnostic correct rate of 72.81% and an overall F1 score of 69.45%. Random forest and XGBoost followed closely, with correct rates of 67.35% and 70.11% and F1 scores of 64.90% and 66.91%, respectively. Linear SVM performed the worst, proving entirely unsuitable for the diagnostic task. Due to its superior performance, CNN was adopted as the classification algorithm in subsequent experiments.

4.3 Optimization analysis of generation strategy

As previously discussed, adding noise to virtual samples can increase data information entropy, thereby enhancing the performance of the trained diagnostic model. This experiment aimed to identify the optimal sample generation strategy by comparing seven noise addition strategies of variable intensity and 20 diffusion function design strategies.

Prompt

For the following question:
Consider the following scenario: A chiller with a rated cooling capacity of 316 kW uses R134A as the refrigerant, and both the condenser and evaporator are configured …

There are two existing answers, but both contain errors:
Answer 1: …
Answer 2: …
Please analyze both answers comprehensively, point out the mistakes, and provide a more accurate answer.

Output of DeepSeek R1

Corrected answer

1. Condenser pressure:
- **Change:** Non-condensable gases (e.g., nitrogen) increase total system pressure by occupying condenser volume and reducing heat transfer efficiency. This elevates the refrigerant's condensing temperature, raising its saturation pressure.
- **Deviation range:** 1.5–3.0 bar increase.
- **Average deviation:** About 2 bar (proportional to non-condensable concentration).

2. Discharge superheat:
- **Change:** Higher condensing pressure forces the compressor to operate at a larger compression ratio, increasing discharge temperature. The refrigerant's saturation temperature also rises, but the discharge temperature escalates disproportionately.
- **Deviation range:** 8–15 °C increase.
- **Average deviation:** About 11 °C (due to compounded inefficiencies in heat rejection and compressor workload).

Key takeaways
- **Answer 1** overestimates condenser pressure and underestimates discharge superheat due to oversimplified heat transfer assumptions.
- **Answer 2** provides closer ranges but requires clearer linkage between non-condensable gases, condensing temperature, and compressor work.
- The corrected ranges align with R134A's thermodynamic properties and water-cooled system behavior under typical fault conditions.

Fig. 7 Verification of multi-model adversarial prompts in LLM

Table 4 Knowledge set of chiller fault symptoms

Type	T_{EI-EO} (°C)		T_{CO-CI} (°C)		P_E (bar)		P_C (bar)		T_{sub} (°C)		T_{suc} (°C)		T_{dis} (°C)	
	Range	Ave.	Range	Ave.	Range	Ave.	Range	Ave.	Range	Ave.	Range	Ave.	Range	Ave.
cf	(0.5, 2.0)	1.25	(1, 3)	2.00	(−0.2, −0.8)	−0.50	(1, 3)	2.00	(−2, −4)	−3.0	–	–	–	–
fwc	(−0.5, 2.5)	1.50	(1, 4)	2.50	(−0.2, −1.0)	−0.60	(1, 4)	2.50	(−1, −3)	−2.0	–	–	–	–
fwe	(1, 4)	2.50	(−0.5, −2.0)	−1.25	(−0.5, −2.0)	−1.25	(−0.5, −2.0)	−1.25	–	–	(2, 8)	5	(2, 10)	6.0
nc	–	–	(0.5, 2.5)	1.50	(−0.1, −0.5)	−0.30	(1.5, 3.0)	2.00	(−1, −5)	−3.0	–	–	(2, 10)	6.0
rl	–	–	(−0.3, −1.5)	−0.90	(−0.2, −0.8)	−0.50	(−0.5, −2.0)	−1.25	(−1, −4)	−2.5	(2, 8)	5	(5, 20)	12.5
ro	–	–	(0.5, 2.0)	1.25	(0.2, 0.8)	0.50	(1, 4)	2.50	(2, 8)	5.0	(−2, −8)	−5	(−2, −10)	−6.0

From Fig. 9, it is evident that virtual samples generated with 5% and 10% noise addition intensity had the highest quality, with 5% showing slightly better average performance (AVG) across Level 1 to Level 4 tests. Several trends can be observed from the specific metric analysis. Firstly, as fault severity increases, the advantage of the noise addition strategy decreases. In the Level 4 test set (the most severe faults), the diagnostic accuracy without noise addition ranked the third among the seven comparisons. Secondly, the effect of noise addition on sample quality improvement initially increased and then decreased with the increase in noise intensity. Additionally, as fault severity increased, the optimal noise addition intensity decreased. In summary, the optimal noise addition intensity was 5% Gaussian noise, indicating that if a parameter deviates by an average of 5 °C, the distribution of this Gaussian noise is centered around $N(0.00, 0.25)$.

As shown in Fig. 10, the quality of virtual samples was validated under 20 different diffusion functions with noise addition strategies of 0%, 5%, and 10%. The correct rate and F1 score showed similar trends. When the upward diffusion distribution remained constant, a smaller n_{lower} value for downward diffusion (larger standard deviation of the function) resulted in higher quality virtual samples. This indicates that expanding the downward sampling range and increasing the probability of sampling small deviation values improves sample generation quality. Conversely, when the downward diffusion distribution remained constant, a larger n_{upper} value for upward diffusion (smaller standard deviation of the function) resulted in higher quality virtual samples. This indicates that reducing the upward sampling range and decreasing the probability of sampling large deviation values enhances sample generation quality. The final generation strategy selected

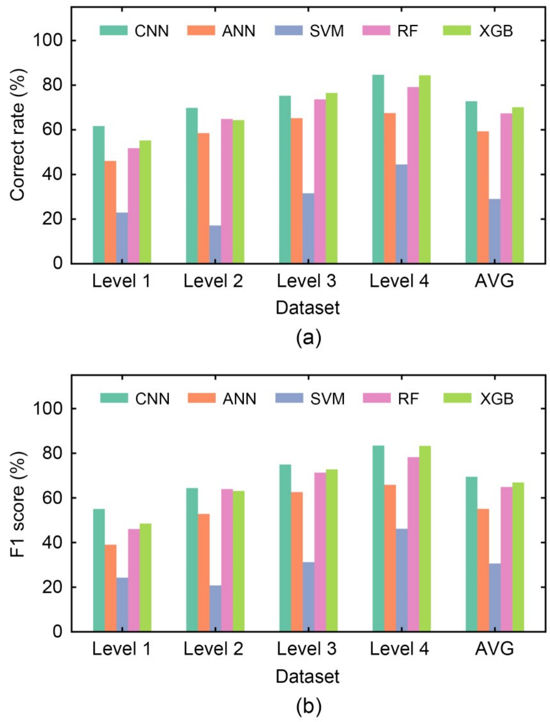

Fig. 8 Comparative analysis of different classifiers: (a) correct rate evaluation; (b) F1 score evaluation. References to color refer to the online version of this figure

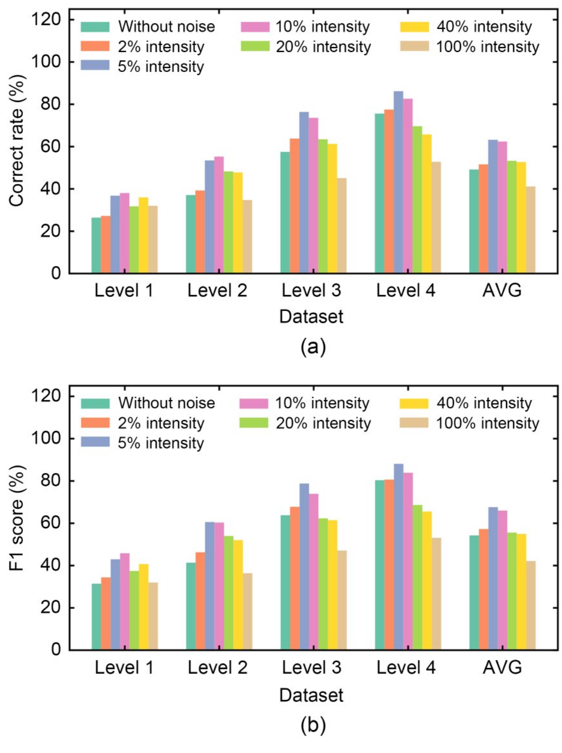

Fig. 9 Comparative analysis of different noise addition intensities: (a) correct rate evaluation; (b) F1 score evaluation. References to color refer to the online version of this figure

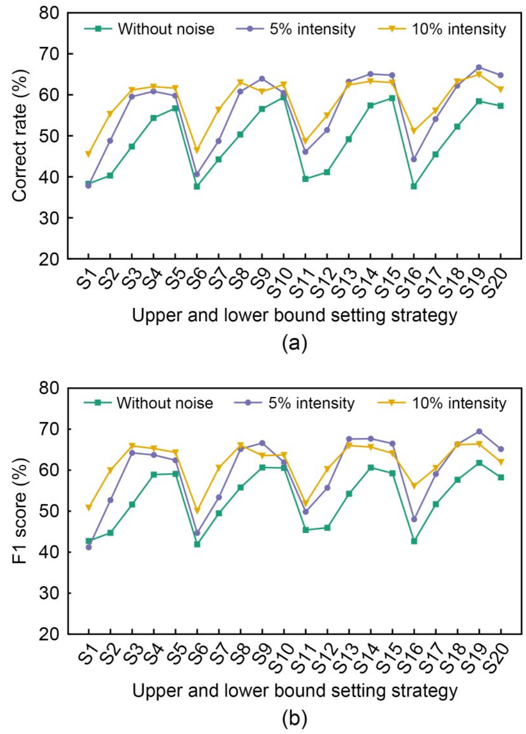

Fig. 10 Comparative analysis of different bound setting strategies for MTD: (a) correct rate evaluation; (b) F1 score evaluation

was the 3/1.5, with an upper boundary of $X_{\text{upper}}=3\sigma_{\text{upper}}$ and a lower boundary of $X_{\text{lower}}=1.5\sigma_{\text{lower}}$.

Since the data values involved in different virtual sample generation scenarios did not vary significantly (there were no differences in magnitude), the hyperparameter optimization values obtained in this section are generally applicable.

4.4 Diagnostic analysis based on optimal strategy

The diagnostic model was trained using virtual samples and evaluated through detailed experimental analysis on a real dataset (Table 5). Overall, the less severe the fault, the harder it was to diagnose accurately. Specifically, considering the average metrics across all fault categories, the accuracy, recall, and F1 score on the Level 1 test set were 61.68%, 49.85%, and 55.03%, respectively. In contrast, on the Level 4 test set, the accuracy, recall, and F1 score were 84.66%, 82.27%, and 83.44%, respectively.

Furthermore, when examining specific faults, the diagnostic model performed best on fwc, fwe, and nc faults, but poorly on cf faults, likely due to the relatively subtle symptoms of cf faults. Additionally, the

Table 5 Detailed evaluation of the optimal generation strategy

Type	Severity	Accuracy (%)	Recall (%)	F1 score (%)
Normal	In Level 1 test dataset	28.66	94.29	43.91
	In Level 2 test dataset	33.92	94.29	49.84
	In Level 3 test dataset	56.84	94.29	70.76
	In Level 4 test dataset	71.13	94.29	80.87
cf	12% reduction in tubes	24.18	4.74	7.54
	20% reduction in tubes	71.03	16.20	25.15
	30% reduction in tubes	51.31	4.94	8.75
	45% reduction in tubes	86.25	28.46	41.53
fwc	10% reduction in flow	100.00	57.83	72.87
	20% reduction in flow	100.00	83.74	90.30
	30% reduction in flow	100.00	99.94	99.97
	40% reduction in flow	99.60	99.40	99.50
fwe	10% reduction in flow	80.24	61.77	67.96
	20% reduction in flow	84.07	86.77	84.77
	30% reduction in flow	96.92	79.80	86.49
	40% reduction in flow	99.52	91.60	94.89
nc	1% by volume nitrogen	90.32	79.43	83.29
	2% by volume nitrogen	83.32	76.37	79.26
	3% by volume nitrogen	80.29	71.69	74.92
	4% by volume nitrogen	85.42	69.80	76.12
rl	10% reduction in charge	70.44	11.51	18.59
	20% reduction in charge	61.84	11.83	19.36
	30% reduction in charge	66.28	82.77	73.32
	40% reduction in charge	79.68	99.97	88.37
ro	10% increase in charge	37.92	39.37	38.41
	20% increase in charge	54.55	49.40	51.43
	30% increase in charge	73.79	90.71	74.95
	40% increase in charge	71.01	92.40	83.44
All	Mean of Level 1	61.68	49.85	55.03
	Mean of Level 2	69.82	59.80	64.39
	Mean of Level 3	75.06	74.88	74.95
	Mean of Level 4	84.66	82.27	83.44
	Mean of Levels 1–4	72.80	66.70	69.45

diagnostic performance on the normal category was very poor, with low accuracy but high recall. This indicates that normal conditions were well-identified, but there was a high rate of faults being misclassified as normal. This is particularly evident in the Level 1 test set, suggesting that the less severe the fault, the more likely it is to be misclassified as normal.

4.5 Comparative validation with supervised learning on real samples

In terms of overall diagnostic performance, the model trained with virtual samples achieved an accuracy of 72.80% and an F1 score of 69.45%, which were close to the accuracy of 73.23% and F1 score of 70.81% obtained from the information-complete dataset (Level 1 to Level 4). This performance was significantly higher than that achieved by training with single severity levels. Training with single severity levels resulted in diagnostic model accuracies of 53.71%, 62.41%, 66.69%, and 66.67%, with corresponding F1 scores of 53.65%, 61.91%, 63.80%, and 63.10%. The proposed method improved accuracy by 19.09%, 10.39%, 6.11%, and 6.13%, and the F1 score by 15.80%, 7.54%, 5.65%, and 6.35%, compared to training with individual datasets. On the four test subsets, Levels 1–4, the models trained on their respective datasets showed significantly better performance, with accuracies of 63.10%, 78.25%, 88.64%, and 88.72%, and F1 scores of 64.42%, 75.15%, 85.97%, and 87.62%, respectively. The model trained with virtual samples nearly achieved the second-best performance across all subsets, with accuracies of 61.68%, 69.82%, 75.06%, and 84.66%, and F1 scores of 55.03%, 64.39%, 74.95%, and 83.44%, respectively. These results indicate that the diagnostic model trained with virtual samples performs well in both diagnostic accuracy and generalization, demonstrating that the proposed virtual sample generation method effectively enhances sample information completeness and addresses the issue of data-driven model training when labeled samples are insufficient (Fig. 11).

4.6 Uncertainty analysis of the proposed method

As shown in Fig. 12, when the knowledge vector output by LLMs varied between 50% and 200%, the diagnostic model's accuracy was indeed affected to some extent. After 50 validation tests across five sample sets, the average diagnostic correct rates were 55.54%, 64.12%, 69.89%, 74.81%, and 66.08%, respectively. Further analysis of their standard deviations, which were 1.01, 1.94, 1.89, 2.01, and 1.45, indicates relatively small fluctuations. Overall, while LLM-generated knowledge did influence the diagnostic outcome, its impact remained moderate and within an interpretable range.

We further examined the scenario where parameter deviations reversed direction, as shown in Fig. 13,

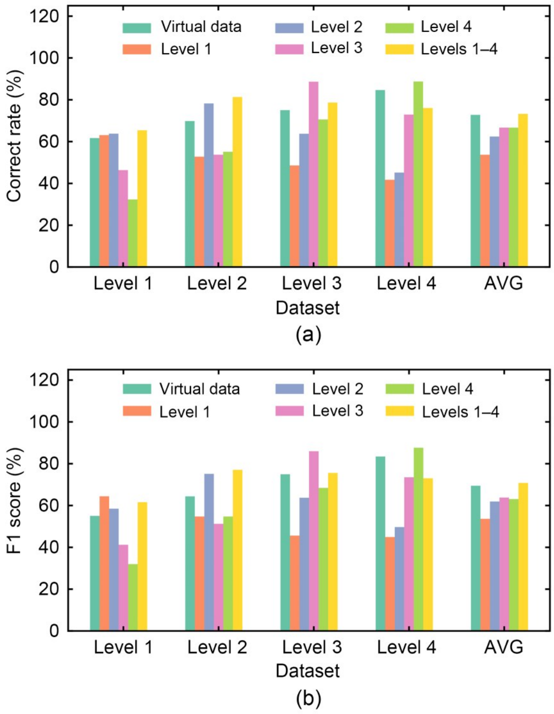

Fig. 11 Comparative analysis with supervised learning: (a) correct rate evaluation; (b) F1 score evaluation. References to color refer to the online version of this figure

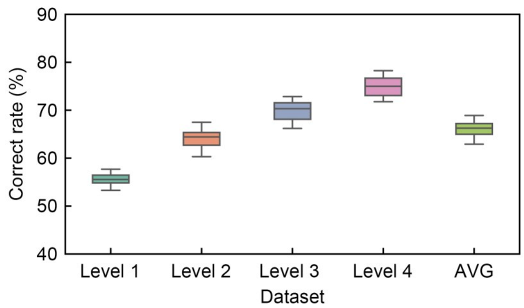

Fig. 12 Uncertainty analysis of the impact of LLM output knowledge on diagnostic accuracy. References to color refer to the online version of this figure

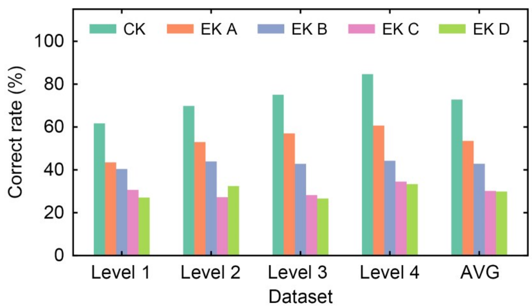

Fig. 13 Comparative analysis of the impact of LLM output erroneous knowledge on diagnostic accuracy. References to color refer to the online version of this figure

with CK denoting correct knowledge and EK denoting error knowledge. EK A–EK D represent cases where 5%, 10%, 15%, and 25% of the shifted knowledge is incorrect (with the shift direction being opposite). Even a 10% misestimation of parameter trends had a substantial impact, reducing the integrated diagnostic accuracy from 72.80% to 53.53% (a decline of 19.27%). When parameter reversal reached 25%, the accuracy dropped to 29.87%, a decrease of 42.93%. These results suggest that misjudging the direction of parameter changes leads to significant losses and severely undermines diagnostic quality.

In summary, while the accuracy of LLM-derived knowledge does affect diagnostic performance, errors in determining parameter change trends pose a more serious threat. Fortunately, our earlier experiments indicate that most LLM responses do not exhibit such misjudgments of parameter trends.

5 Conclusions

In this paper, we propose a knowledge-guided virtual sample generation method. Firstly, the domain knowledge required for sample generation was acquired using ChatGPT-4o. This domain knowledge was then used to optimize the boundaries of the trend diffusion function. Virtual samples were generated by sampling within the diffusion function space using the Monte Carlo method. Finally, a noise-adding strategy was used to enhance the information entropy of the samples, thereby improving the robustness and generalization of the diagnostic model trained with these samples. The main conclusions were as follows:

(1) A new approach to acquiring domain knowledge using ChatGPT-4o is proposed. A basic framework for the domain knowledge required for sample generation was established, and a guided general model dialogue process was designed to obtain more reliable and effective domain knowledge.

(2) A KG-MTD strategy is introduced. Domain knowledge was used to design diffusion function boundaries, replacing the small sample boundary estimation method in the original MTD. Monte Carlo sampling was used in the diffusion space to generate a complete virtual sample set without real samples.

(3) A method to enhance sample information entropy based on noise addition is proposed. By adding Gaussian noise of a certain intensity to the virtual

samples, the information entropy of the sample set was increased, thereby enhancing the robustness and generalization of the trained neural network.

(4) Tests on an open-source building air conditioning dataset showed that the proposed method can generate high-quality and complete virtual sample sets. The comprehensive accuracy of the trained diagnostic model reached 72.80%, which is close to the accuracy of models trained with real samples, with significant advantages in generalization.

(5) Finally, an analysis of the uncertainty in LLM-based diagnosis showed that, when parameter trends were correctly identified, fluctuations in deviation exerted minimal impact on diagnostic outcomes. However, misjudging parameter trends led to a significant drop in accuracy: even a 5% misestimation of parameters reduced accuracy by 19.27%.

In summary, the knowledge-guided virtual sample generation method proposed in this paper can generate information-complete virtual sample sets without small real samples, addressing the issue of zero-sample fault diagnosis in building air conditioning systems. The trained diagnostic model also has superior generalization. This method can be extended to other thermal system fault diagnosis fields, showing great potential.

Generative AI usage declaration

During this work, the authors used generative AI models, DeepSeek, ChatGPT o3-mini-high, and ChatGPT 4o, to generate expert knowledge for fault diagnosis. All AI-produced contents were reviewed and validated by domain experts, and subsequently edited by the authors, who take full responsibility for the final content in compliance with journal policies on ethical AI use.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62306281), the Natural Science Foundation of Zhejiang Province (Nos. LQ23E060006 and LTGG24E050005), and the Key Research Plan of Jiaxing City (No. 2024BZ20016).

Author contributions

Zhe SUN: conceptualization, methodology, software, investigation, formal analysis, writing-review and editing; Qiwei YAO: software, methodology; Ling SHI: software, methodology; Huaqiang JIN: supervision; Yingjie XU: formal analysis; Peng YANG: software; Han XIAO: software; Dongyu CHEN: methodology; Panpan ZHAO: methodology; Xi SHEN: supervision, funding acquisition, writing-review.

Conflict of interest

Zhe SUN, Qiwei YAO, Ling SHI, Huaqiang JIN, Yingjie XU, Peng YANG, Han XIAO, Dongyu CHEN, Panpan ZHAO, and Xi SHEN declare that they have no conflict of interest.

References

Bishop CM, 1995. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108-116. https://doi.org/10.1162/neco.1995.7.1.108

Chen KL, Wang ZW, Gu XW, et al., 2021. Multicondition operation fault detection for chillers based on global density-weighted support vector data description. *Applied Soft Computing*, 112:107795. https://doi.org/10.1016/j.asoc.2021.107795

Chen ZL, O'Neill Z, Wen J, et al., 2023. A review of data-driven fault detection and diagnostics for building HVAC systems. *Applied Energy*, 339:121030. https://doi.org/10.1016/j.apenergy.2023.121030

Comstock MC, Braun JE, Groll EA, 2001. The sensitivity of chiller performance to common faults. *HVAC&R Research*, 7(3):263-279. https://doi.org/10.1080/10789669.2001.10391274

Du M, Ren FF, Min R, et al., 2024. Detecting non-uniform structures in oil-in-water bubbly flow experiments. *Physica A: Statistical Mechanics and Its Applications*, 637:129602. https://doi.org/10.1016/j.physa.2024.129602

Forth K, Borrmann A, 2024. Semantic enrichment for BIM-based building energy performance simulations using semantic textual similarity and fine-tuning multilingual LLM. *Journal of Building Engineering*, 95:110312. https://doi.org/10.1016/j.jobe.2024.110312

Frank SJ, Frank AM, 2020. Salient slices: improved neural network training and performance with image entropy. *Neural Computation*, 32(6):1222-1237. https://doi.org/10.1162/neco_a_01282

Gappmair W, 1999. Claude E. Shannon: the 50th anniversary of information theory. *IEEE Communications Magazine*, 37(4):102-105. https://doi.org/10.1109/35.755458

Guo YB, Liu YX, Zhang Z, et al., 2024. Research on fault detection and diagnosis of carbon dioxide heat pump systems in buildings based on transfer learning. *Journal of Building Engineering*, 85:108774. https://doi.org/10.1016/j.jobe.2024.108774

Holmstrom L, Koistinen P, 1992. Using additive noise in back-propagation training. *IEEE Transactions on Neural Networks*, 3(1):24-38. https://doi.org/10.1109/72.105415

Huang CF, 1997. Principle of information diffusion. *Fuzzy Sets and Systems*, 91(1):69-90. https://doi.org/10.1016/S0165-0114(96)00257-6

Huang CF, Moraga C, 2004. A diffusion-neural-network for learning from small samples. *International Journal of Approximate Reasoning*, 35(2):137-161. https://doi.org/10.1016/j.ijar.2003.06.001

Jiang G, Ma ZH, Zhang L, et al., 2024. EPlus-LLM: a large language model-based computing platform for automated

building energy modeling. *Applied Energy*, 367:123431.
https://doi.org/10.1016/j.apenergy.2024.123431

Khamis N, Selamat H, Ismail FS, 2022. Improved optimization parameters prediction using the modified mega trend diffusion function for a small dataset problem. *Knowledge and Information Systems*, 64(11):3129-3149.
https://doi.org/10.1007/s10115-022-01727-z

Li DC, Wu CS, Tsai TI, et al., 2007. Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. *Computers & Operations Research*, 34(4):966-982.
https://doi.org/10.1016/j.cor.2005.05.019

Li GN, Yao Q, Fan C, et al., 2021. An explainable one-dimensional convolutional neural networks based fault diagnosis method for building heating, ventilation and air conditioning systems. *Building and Environment*, 203:108057.
https://doi.org/10.1016/j.buildenv.2021.108057

Li GN, Chen L, Liu JY, et al., 2023a. Comparative study on deep transfer learning strategies for cross-system and cross-operation-condition building energy systems fault diagnosis. *Energy*, 263:125943.
https://doi.org/10.1016/j.energy.2022.125943

Li GN, Xiong JH, Tang R, et al., 2023b. In-situ sensor calibration for building HVAC systems with limited information using general regression improved Bayesian inference. *Building and Environment*, 234:110161.
https://doi.org/10.1016/j.buildenv.2023.110161

Li L, Li QH, Ni YS, et al., 2024a. Critical penetrating vibration evolution behaviors of the gas-liquid coupled vortex flow. *Energy*, 292:130236.
https://doi.org/10.1016/j.energy.2024.130236

Li L, Xu P, Xu WX, et al., 2024b. Multi-field coupling vibration patterns of the multiphase sink vortex and distortion recognition method. *Mechanical Systems and Signal Processing*, 219:111624.
https://doi.org/10.1016/j.ymssp.2024.111624

Li L, Xu P, Li QH, et al., 2025a. A coupled LBM-LES-DEM particle flow modeling for microfluidic chip and ultrasonic-based particle aggregation control method. *Applied Mathematical Modelling*, 143:116025.
https://doi.org/10.1016/j.apm.2025.116025

Li L, Xu P, Li QH, et al., 2025b. Multi-field coupling particle flow dynamic behaviors of the microreactor and ultrasonic control method. *Powder Technology*, 454:120731.
https://doi.org/10.1016/j.powtec.2025.120731

Li TT, Zhou YZ, Zhao Y, et al., 2022a. A hierarchical object oriented Bayesian network-based fault diagnosis method for building energy systems. *Applied Energy*, 306:118088.
https://doi.org/10.1016/j.apenergy.2021.118088

Li TT, Zhao Y, Zhang CB, et al., 2022b. A semantic model-based fault detection approach for building energy systems. *Building and Environment*, 207:108548.
https://doi.org/10.1016/j.buildenv.2021.108548

Li Z, Wang CY, Li L, et al., 2024. Numerical investigation of mesoscale multiphase mass transport mechanism in fibrous porous media. *Engineering Applications of Computational Fluid Mechanics*, 18(1):2363246.
https://doi.org/10.1080/19942060.2024.2363246

Liu JY, Li X, Zhang Q, et al., 2023. An efficient sensor and thermal coupling fault diagnosis methodology for building energy systems. *Energy and Buildings*, 296:113367.
https://doi.org/10.1016/j.enbuild.2023.113367

Lu J, Zhang CB, Li JY, et al., 2022. Graph convolutional networks-based method for estimating design loads of complex buildings in the preliminary design stage. *Applied Energy*, 322:119478.
https://doi.org/10.1016/j.apenergy.2022.119478

Lu J, Tian XN, Zhang CB, et al., 2025. Evaluation of large language models (LLMs) on the mastery of knowledge and skills in the heating, ventilation and air conditioning (HVAC) industry. *Energy and Built Environment*, 6(5):875-892.
https://doi.org/10.1016/j.enbenv.2024.03.010

Lu JG, Li DD, 2013. Bias correction in a small sample from big data. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2658-2663.
https://doi.org/10.1109/TKDE.2012.220

Niyogi P, Girosi F, Poggio T, 1998. Incorporating prior information in machine learning by creating virtual examples. *Proceedings of the IEEE*, 86(11):2196-2209.
https://doi.org/10.1109/5.726787

Qi GJ, Luo JB, 2022. Small data challenges in big data era: a survey of recent progress on unsupervised and semi-supervised methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2168-2187.
https://doi.org/10.1109/TPAMI.2020.3031898

Sanchez FAS, Khambampati AK, Kim KY, 2023. Generative adversarial network model for two-phase flow imaging by electrical impedance tomography. *IEEE Transactions on Instrumentation and Measurement*, 72:4507212.
https://doi.org/10.1109/TIM.2023.3315421

Sivakumar J, Ramamurthy K, Radhakrishnan M, et al., 2022. Synthetic sampling from small datasets: a modified mega-trend diffusion approach using *k*-nearest neighbors. *Knowledge-Based Systems*, 236:107687.
https://doi.org/10.1016/j.knosys.2021.107687

Sun Z, Yao QW, 2024. Self-correction method for sensor faulty heat pump system based on machine learning. *Results in Engineering*, 22:102170.
https://doi.org/10.1016/j.rineng.2024.102170

Sun Z, Jin HQ, Gu JP, et al., 2019. Gradual fault early stage diagnosis for air source heat pump system using deep learning techniques. *International Journal of Refrigeration*, 107:63-72.
https://doi.org/10.1016/j.ijrefrig.2019.07.020

Sun Z, Jin HQ, Gu JP, et al., 2020. Studies on the online intelligent diagnosis method of undercharging sub-health air source heat pump water heater. *Applied Thermal Engineering*, 169:114957.
https://doi.org/10.1016/j.applthermaleng.2020.114957

Sun Z, Jin HQ, Xu YJ, et al., 2022. Severity-insensitive fault diagnosis method for heat pump systems based on improved benchmark model and data scaling strategy. *Energy and Buildings*, 256:111733.
https://doi.org/10.1016/j.enbuild.2021.111733

Sun Z, Yao QW, Jin HQ, et al., 2024. A novel in-situ sensor calibration method for building thermal systems based on virtual samples and autoencoder. *Energy*, 297:131314. https://doi.org/10.1016/j.energy.2024.131314

Tan YF, Ni YS, Xu WX, et al., 2023. Key technologies and development trends of the soft abrasive flow finishing method. *Journal of Zhejiang University-SCIENCE A*, 24(12):1043-1064. https://doi.org/10.1631/jzus.A2300038

Tian Y, Dong QY, Tian JD, et al., 2023. Capacity estimation of lithium-ion batteries based on optimized charging voltage section and virtual sample generation. *Applied Energy*, 332:120516. https://doi.org/10.1016/j.apenergy.2022.120516

Wang ZW, Wang ZW, He SW, et al., 2017. Fault detection and diagnosis of chillers using Bayesian network merged distance rejection and multi-source non-sensor information. *Applied Energy*, 188:200-214. https://doi.org/10.1016/j.apenergy.2016.11.130

Wang ZW, Wang L, Tan YY, et al., 2021. Fault diagnosis using fused reference model and Bayesian network for building energy systems. *Journal of Building Engineering*, 34:101957. https://doi.org/10.1016/j.jobe.2020.101957

Wei CH, Ooka R, 2023. Indoor airflow field reconstruction using physics-informed neural network. *Building and Environment*, 242:110563. https://doi.org/10.1016/j.buildenv.2023.110563

Whittaker T, Janik RA, Oz Y, 2024. Turbulence scaling from deep learning diffusion generative models. *Journal of Computational Physics*, 514:113239. https://doi.org/10.1016/j.jcp.2024.113239

Wu JF, Li L, Yin ZC, et al., 2024. Mass transfer mechanism of multiphase shear flows and interphase optimization solving method. *Energy*, 292:130475. https://doi.org/10.1016/j.energy.2024.130475

Xu ML, Yoon S, Fuentes A, et al., 2023. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, 137:109347. https://doi.org/10.1016/j.patcog.2023.109347

Xu P, Li QH, Wang CY, et al., 2025. Interlayer healing mechanism of multipath deposition 3D printing models and interlayer strength regulation method. *Journal of Manufacturing Processes*, 141:1031-1047. https://doi.org/10.1016/j.jmapro.2025.03.062

Xu YJ, Wang JF, Shen X, et al., 2023. Thermodynamic analyses and performance improvement on a novel cascade-coupling-heating heat pump system for high efficiency hot water production. *Energy Conversion and Management*, 293:117448. https://doi.org/10.1016/j.enconman.2023.117448

Yu FW, Li GN, Chen HX, et al., 2018. A VRF charge fault diagnosis method based on expert modification C5.0 decision tree. *International Journal of Refrigeration*, 92:106-112. https://doi.org/10.1016/j.ijrefrig.2018.05.034

Yu XR, He YL, Xu Y, et al., 2019. A mega-trend-diffusion and monte carlo based virtual sample generation method for small sample size problem. *Journal of Physics: Conference Series*, 1325:012079. https://doi.org/10.1088/1742-6596/1325/1/012079

Zhang CB, Tian XN, Zhao Y, et al., 2022. Causal discovery-based external attention in neural networks for accurate and reliable fault detection and diagnosis of building energy systems. *Building and Environment*, 222:109357. https://doi.org/10.1016/j.buildenv.2022.109357

Zhang CB, Zhang J, Zhao Y, et al., 2024a. Automated data mining framework for building energy conservation aided by generative pre-trained transformers (GPT). *Energy and Buildings*, 305:113877. https://doi.org/10.1016/j.enbuild.2023.113877

Zhang CB, Lu J, Huang JH, et al., 2024b. End-to-end data-driven modeling framework for automated and trustworthy short-term building energy load forecasting. *Building Simulation*, 17(8):1419-1437. https://doi.org/10.1007/s12273-024-1149-y

Zhang J, Zhang CB, Lu J, et al., 2025. Domain-specific large language models for fault diagnosis of heating, ventilation, and air conditioning systems by labeled-data-supervised fine-tuning. *Applied Energy*, 377:124378. https://doi.org/10.1016/j.apenergy.2024.124378

Zhao Y, Li TT, Zhang XJ, et al., 2019. Artificial intelligence-based fault detection and diagnosis methods for building energy systems: advantages, challenges and the future. *Renewable and Sustainable Energy Reviews*, 109:85-101. https://doi.org/10.1016/j.rser.2019.04.021

Zhou M, Liu TY, Li Y, et al., 2019. Toward understanding the importance of noise in training neural networks. Proceedings of the 36th International Conference on Machine Learning, p.7594-7602.